Creating the Daily Mail Historical Archive

Seth Cayley, Publisher for Media History at Cengage Learning

It is self-evident that digital archives have transformed the landscape of historical research, especially of the eighteenth and nineteenth centuries, on which the majority of these resources are concentrated. The sheer digital accessibility of rare material, which in physical form can be locked away in the special collections of a single library, has helped to democratize the study of the past. In tandem with this, many digital archives permit word-searching of the content, opening up the material to a form of textual interrogation not permissible in print.

In this gold-rush to digital, there is a danger of taking it all for granted. There is a common misconception that digitizing the archive of a newspaper is simply a case of scanning the pages and putting them on the internet. In this regard, digital archives are a victim of their own success; the best are so simple to use that they look equally simple to put together. Yet behind each digital newspaper archive is a mammoth project involving editorial selection, content processing conundrums and a wide variety of bespoke technical decisions.
Jim Mussell has rightly argued that 'these resources actually constitute a type of edition … Users must be able to analyze how a resource has been put together if they are to understand how the digital representation differs from whatever it republishes.'[1] By presenting the methodology used in creating the Daily Mail Historical Archive, my aim here is to bring to users' attention the history of this digital edition, the transformations that have taken place to the content, and an appreciation of the scale of the whole initiative.

Sourcing the Material

The first headache encountered by any project to digitize an historic newspaper is which edition do we use? To those unfamiliar with media history it may come as a surprise to learn that there is no single edition of a daily newspaper. Historically, most papers have published – and continue to publish – multiple editions in one day, including late editions, regional editions and weekly editions. While the content will be broadly the same between the editions, there will be some differences in the selection of news stories, advertising, and even seemingly trivial details such as the masthead. Which, then, is the 'authoritative' edition to use for digitization? Perhaps in an ideal world we would digitize all editions of the paper, so that researchers could see it in all its forms. Unsurprisingly, the costs of doing so make this a non-starter in most cases; digitizing just one edition of the paper is expensive enough. But there is also the experience of the end user to consider. Does someone searching a newspaper archive really want their results swollen by multiple editions? No doubt there are scholars who would derive benefit from this, but our experience shows that newspaper archives are used by a broad church of the research community, including family historians, science departments, and schools. Keeping the archive straightforward and non-confusing is vital.

For content dating before the 1970s and the heyday of microfilming, most archives are digitized from the versions of the newspaper or periodical that were bound into annual or semi-annual volumes. These 'library editions' were produced by publishers precisely for preservation and collecting purposes. As Laurel Brake points out, these editions normally exclude 'ephemeral, paratextual matter' such as advertising wrappers, supplements, and in some cases even covers.[2] library editions are, by necessity, the versions of newspapers most widely consulted in print, but it is important to note that they are no more definitive than other editions. From the 1970s onwards, most major UK newspapers instigated microfilming programmes to preserve their paper for archival purposes. Although many continued to publish bound library editions, microfilming represented a shift in the preservation process, with loose editions being filmed on a monthly basis. At the same time, most newspapers microfilmed their back issues of the bound library editions, so that they had a complete corpus of the newspaper on microfilm.

In the case of the Daily Mail Historical Archive, it is this in-house microfilm, owned by Associated Newspapers, which we have used as the basis for digitization. On the microfilm, the final London edition of the newspaper is

[1] Jim Mussell, 'Teaching Nineteenth-Century Periodicals Using Digital Resources: Myths and Methods', *Victorian Periodicals Review,* Vol 45, No. 2, pp.201–09.
[2] Laurel Brake, 'The Longevity of "Ephemera": Library editions of nineteenth-century periodicals and newspapers", *Media History*, Volume 18, Issue 1, February 2012, pp. 7–20.

the edition that has been used for filming, so in the majority of cases this is what the user will be viewing. Occasionally, where the final edition of the day was not available, we have used earlier editions.

Using the microfilm led to some problems caused by the medium. There were a number of issues, particularly in the early years of the paper, where the team determined that the filmed images were unfit for purpose; the original source was torn or damaged and should never have been used for filming; or the issues had been filmed from tightly bound volumes, causing excessive curvature and obscured text. In most such cases, we replaced the images using reels from an alternative microfilm edition created separately by the British Library. It is worth stressing that microfilm is not innately an inferior medium for digitization. If the original filming was done well and from well-preserved original copies, even decades-old microfilm can produce surprisingly good digital images.

Working our way through the microfilm threw up some surprising anomalies too, evidence of selection decisions made by earlier editors. During the General Strike of 1926, Fleet Street did not put out any newspapers. On the microfilm, the Daily Mail Continental Edition, published in Paris, plugs the gap. We elected to keep this in the digital archive, rather than have no edition available for the period of the strike.

Where available, supplements have been included. Supplements are defined here as components of the paper that have separate pagination from the main part of the newspaper. In print, these sometimes appear in the middle of the paper. This makes sense in terms of packaging the newspaper up as a physical object, to make it compact and easily foldable. In the digital edition it does not make sense to replicate the exact place in which these supplementary materials appeared, as it renders the page sequences extremely confusing, and sometimes splits articles in two. Users cannot pull out the magazine and other supplements to read the main paper uninterrupted, as they would in the material world. We have therefore placed supplements after the paper.

The Weekend magazine, included as a Saturday supplement from 1992, has been filmed inconsistently, and as a result we do not have a complete run in the digital archive. For cost reasons, it has not been possible to fill the gaps by scanning the missing issues. With regret, the primary focus of this project is the main paper. There are other necessary omissions. There are Scottish and Irish editions of the paper, as well as the Continental Daily Mail and the short-lived weekly US edition published from 1944 to 1946. There was even a Braille edition. Ultimately it has not been feasible to include these. In a related vein, the weekly Mail on Sunday began publication in 1983. Although under the same ownership as its daily sister paper, the Mail on Sunday is a separate newspaper, with its own editorial and journalistic team. Given this, and that its historical legacy is not as long as the Daily Mail, we felt justified in de-scoping it from the project. If there is significant interest, the Mail on Sunday may be added to the archive as an upgrade in the future. Finally, Alfred Harmsworth created no less than 63 ‘dummy’ editions of the Daily Mail before its official launch on 4 May 1896, to test his ideas about layout, visual impact, and balance of content. Only some of these ‘dummy’ editions survive, and they were never released to the public, so they have not been included in what is intended to be an archive of the ‘official’ newspaper.

Why Stop at 2004?

Determining the cut-off date for an archive of any newspaper still in print presents a dilemma. One person’s logical end-date is another researcher’s anguish that XYZ news story falls after that date. Commercial realities dominate the decision-making process here. Although the approved budget for the project was generous – the final costs will be well into seven figures in pounds sterling – it was not unlimited. Every additional year that we opted to include added to the total page count, which in turn added to the content-processing costs. The problem became particularly acute from the 1990s onwards, when newspapers such as the Daily Mail exploded in size. It is noteworthy that the Daily Mail Historical Archive has the largest page count of any single-title newspaper archive that Cengage Learning has created.

Our original intention had been to digitize the first 100 years (1896–1996), but as the project went into pre-production we realized that, by ending four years short of the millennium, we were missing an opportunity to include a complete perspective of the twentieth century.

As it transpired, the main batch of Associated Newspapers microfilm stops in 2004, so the revised cut-off date presented itself, and additional budget was secured. Thereafter, full-text editions exist on various aggregator

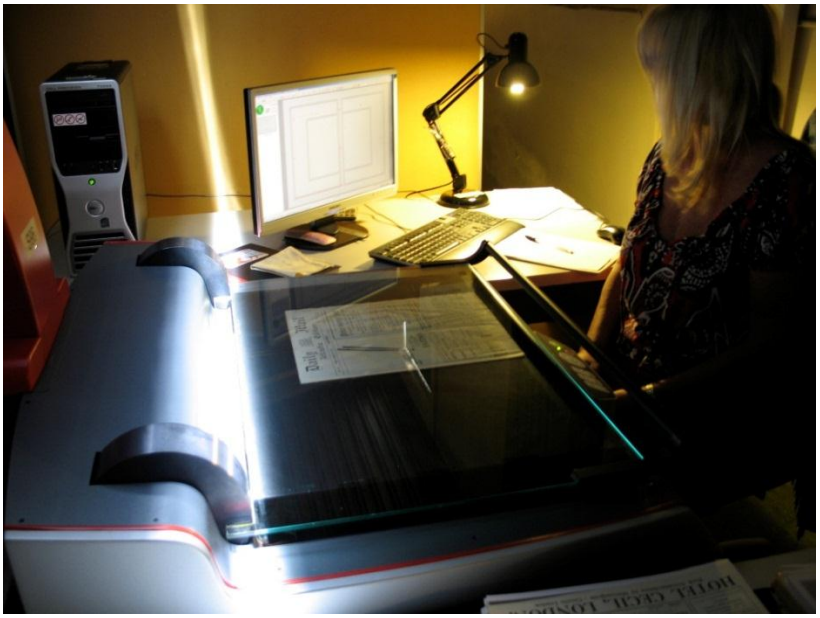
sites, although the user experience is not the same. If demand exists, it may be possible to incorporate the Daily Mail's in-house PDFs of the post-2004 years at a later stage, but they cannot simply be uploaded. As with the digital files generated from the microfilm, they require the generation of XML files (see below) that describe the content structure and allow them to be a functional part of a digital archive.

The Atlantic Edition

Seemingly contradicting our stated policy of not including multiple editions, we have included the Atlantic Edition in the archive (see 'The Daily Mail Atlantic Edition'). Printed at sea, these issues are extremely rare, and even the British Library does not hold copies. The only known set is held by Associated Newspapers itself. Long-neglected, the bound volumes have suffered rodent, water and other storage damage over the years, and are deteriorating fast. Preservation of this forgotten source for the history of the 1920s was therefore deemed by the team as an important task for future generations.

Bound volumes of the Daily Mail Atlantic Edition, showing both wireless page volumes and the complete volume for the A boat in a state of disrepair

The Atlantic Edition was scanned from the physical copies, as no microfilm exists. The bindings were removed to permit flat scanning, and the volumes were subsequently re-bound and put into deep vacuum storage to prevent further decay. Some issues were in poor condition or missing entirely, but unfortunately no replacement copies were available.



Scanning the Atlantic Edition

The organization of the Atlantic Edition presented its own challenges, as the system had been devised by an Associated Newspapers archivist several decades previously, who had taken the secret to the grave; it took time to crack the code.

Upon investigation, we ascertained that the Atlantic Editions were organized as*:*
A boats: the largest Cunard liners (Aquitania, Berengaria, Mauretania)
B boats: the smaller ships
wireless volumes

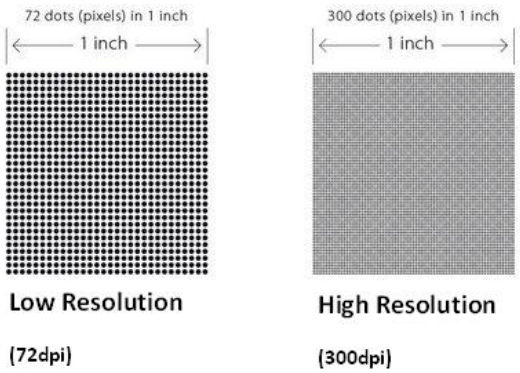
The archivist preserved a complete copy of each issue of the A boat, and a complete copy of each issue for the B boats. How the archivist selected which A boat and B boat to use as the preservation copy is unknown. It may have depended on what issues were to hand. A boats are usually the Berengaria or Aquitania, although there are some issues from the Mauretania.

The wireless volumes are more haphazard, sometimes containing just the wireless pages from a given boat, containing many repeats, and sometimes issues from boats that were sailing at the same time as the A or B boat. These give the impression as volumes where everything else was 'stuffed in'.
Once we realized that many volumes contained repeats, we could have made a justifiable editorial decision to film only the A volumes. However, we realized that some of the B volumes were sailing at times when there was no A ship at sea. We therefore decided to do both the A and B volumes. This means there are some dates when there is more than one Atlantic edition available, allowing for interesting comparisons about how the respective typesetters presented the news. It may also highlight different advertising: passengers on B ships were probably not quite as wealthy as the A passengers.

Of the original 80 volumes (an estimated 300,000 pages), we have scanned 26 (approximately 40,000 pages). Filming all the Atlantic editions was not a realistic prospect – it would have consumed a quarter of the project budget – but by filming the A and B volumes we have at least been able to capture the two whole runs preserved by the original archivist.

Image Capture
The process by which the digital images are created – known as 'image capture' – fundamentally affects what the user sees on screen. Images captured at a very high resolution, whereby significant amounts of detail are recorded per inch of space, will look much more photo-realistic and permit high zooming percentages before appearing blocky. However, this comes at the price of vast file sizes and consequently slow loading times. A balance has to be struck between quality of the image and versatility of use.

Images captured from microfilm were therefore captured at a resolution of 400dpi (dots per inch), in line with standards provided by the Library of Congress for digitization projects. For medieval manuscripts a higher resolution (perhaps as much as 1200dpi) could be suitable, in order to permit close viewing of intricate details, but with newspapers a resolution of 400dpi provides a reasonable balance between readability and manageable file size.



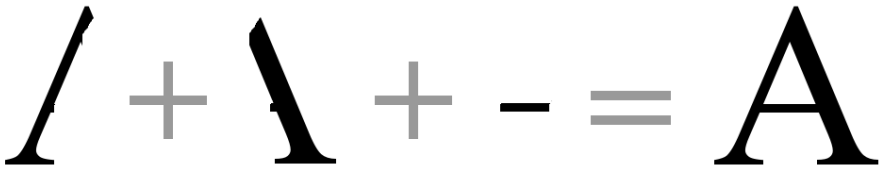
Comparison of 72dpi and 300dpi, showing the fineness of detail that the higher resolution provides. The Daily Mail Historical Archive uses an even higher resolution of 400dpi. Image reproduced with the kind permission of www.zenfolio.com

A decision was also required about whether to capture the Daily Mail images as bitonal (i.e. simple black and white) or as greyscale (which permits subtlety of tone and different shades). Each has their pros and cons. Bitonal can make text very stark and clear against a white background, and is regularly used for historic newspapers. However, it renders illustrations and photographs very poorly. By contrast, greyscale handles photographs and illustrations much better, but also picks up the background grain of the paper, meaning that the images have a grey 'noisy' background. As the Daily Mail becomes increasingly illustrated during the course of the twentiethth century – by the end actually printing articles within giant photos – we needed to use greyscale for the later content. We therefore opted for a hybrid solution, where the images would be bitonal up to 1962, when colour and true halftone photographs start appearing more regularly, and greyscale thereafter. Users will notice a contrast in the way the images appear either side of this date.

The Daily Mail switched from broadsheet to tabloid format on 3 May 1971. It was noticeable that the file sizes of the broadsheet years were significantly higher than those after the move to tabloid format, especially the greyscale images. In order to achieve manageable file sizes, the dimensions of the pre-tabloid Daily Mail were reduced by 50%, so the broadsheet issues in the archive will appear to be the same physical size as the tabloid years. In many ways, this makes the paper much easier to read on screen – few people have a broadsheet-sized monitor – but it is worth highlighting this as a transformation brought about by the digital edition.
The Atlantic Edition was filmed at 400dpi, and captured as bitonal. The originals are tabloid-sized, and there was no reduction in size for the digital edition. The special issues were captured at 400dpi in full colour at their original size.

OCR
Scanned pages of a newspaper are simply a form of photograph – a picture of the text – and consequently of limited use in of themselves. Without data that supports those images, the scanned pages are not searchable or discoverable in a digital environment. The creation of such data is the key component of any digital archive project. It powers the functionality that allows users to search, retrieve and browse the hundreds of thousands of pages.

To render the text on a scanned newspaper page searchable, we put it through a process known as optical character recognition (OCR). The text produced by the OCR process is what is actually being checked when a user enters a search term. OCR software analyses the light and dark areas of the scanned image in order to identify each alphabetic letter and numeric digit. When it recognizes a character, it converts it into regular text.

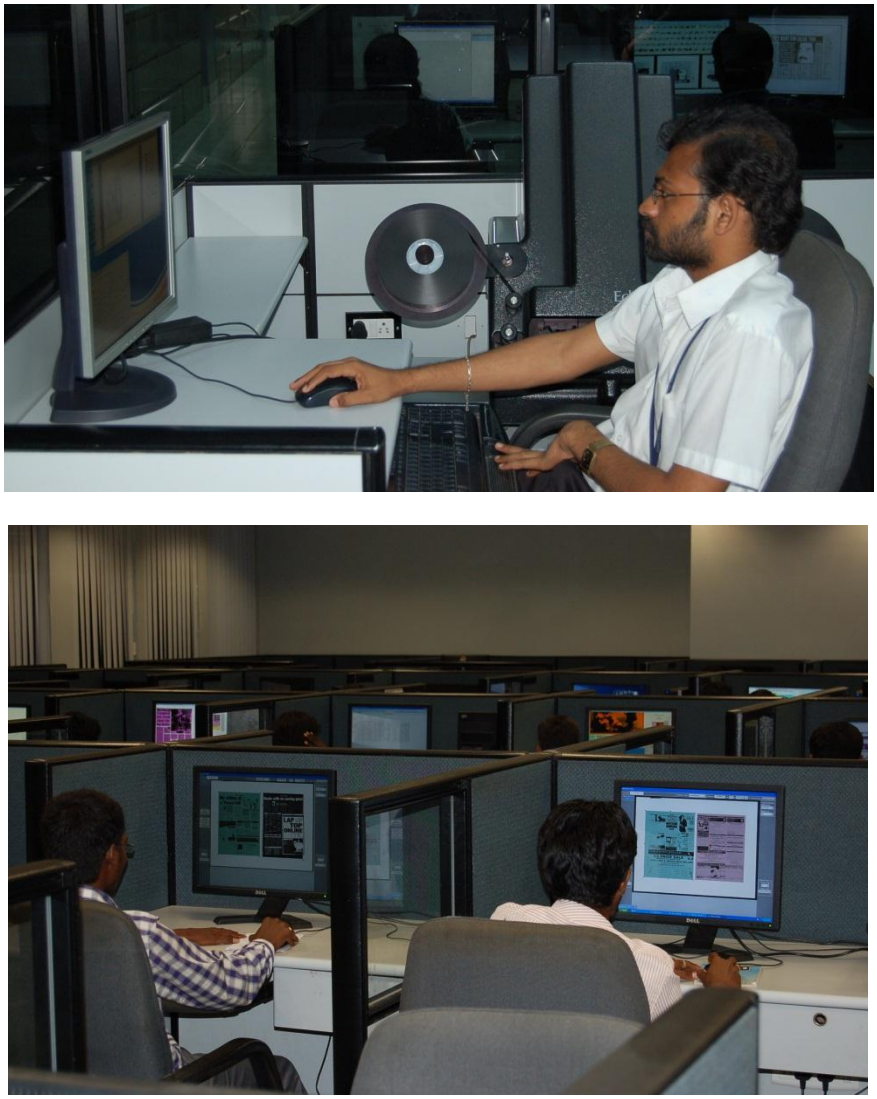


Example of how modern OCR software carries out 'feature detection'.By learning the common features of a letter or character, the OCR software can recognize most letters, whatever the typeface. In this case a capital 'A' will always have two slanting lines and a bridging line in between. Image reproduced with the kind permission of www.explainthatstuff.com

OCR is an imperfect process, and there is a wide array of challenges. The quality of the OCR text usually says more about the condition of the original materials than it does about the performance of the OCR software. Certain types of material are much harder to OCR than others. As a general rule of thumb, older newspapers produce much less satisfactory results than modern ones. This can be because the originals are worn or in a poor condition, or because the text is smudged and difficult to read. Any document that was printed by hand is much more difficult for OCR software to analyze than machine-printed characters. Wartime newspapers

noticeably produce poorer results than adjacent periods of history; they were usually printed on poor-quality, thin paper owing to rationing, which leads to 'bleed-through' of text from the other side of the page.
People often ask 'how accurate is the OCR in your digital archives?' The word 'accuracy' is misleading here, because OCR software works from a *confidence* rating, not true accuracy. The software calculates a confidence level from 0 to 9 for each character it detects, but does not know whether a character has been converted correctly or not. The software can only be confident or not confident that it is correct.

True accuracy, that is, whether a character is actually correct, can only be determined by a human assessing each character manually. This is why it is not possible to correct the OCR of these projects, at least initially. The Daily Mail Historical Archive had a team of over 400 people creating and reviewing the data for the archive, but with over 1.2 million pages to digitize and convert, it was not physically possible to clean up the OCR for every article. At the time of writing, only small-scale digitization projects have a realistic opportunity of being able to produce 100% perfect OCR, and even those projects typically rely on the goodwill and time of unpaid enthusiasts manually correcting the OCR text.



Members of the 400-strong team involved in the creation of digital data. In the top photograph, the operator is scanning the microfilm to digital. In the bottom photograph the team are reviewing the digital images and data

Metadata

While we cannot guarantee perfect OCR text in these projects, we do aim for 99.5% accuracy in the *metadata* we produce. Metadata is the 'who, what, where and when' of a digital object, providing it with key descriptive information that permits it to be organized more easily. If quality of the metadata is high, it becomes much easier to find the specific types of information a researcher is looking for, and to place effective parameters and filters around a search query, such as date ranges limited to specific article types.

Metadata is created at several layers, including publication level, issue level, page level and article level. In the case of a newspaper article in an historical archive, the metadata will include:

- Article title (or first line of text if no formal title is present)
- Author (if known)
- Newspaper title
- Date of publication
- Issue/edition Number of the newspaper
- Page number
- Article category (e.g. Advertising, News, Letter).

Most of this metadata is entered manually, and then verified by two operators working independently, to ensure they agree on the same result. If they disagree, additional opinions are sought.
We ensure that our metadata is consistent across our newspaper archives, which allows a familiar user experience, and permits cross-searching and cross-browsing.

In the case of the Daily Mail, we decided to create additional metadata fields to accommodate the Atlantic Edition. Following extensive checking that involved looking at every single issue, we determined two common threads that were specific to all issues of the Atlantic Edition: name of the ship and the direction of travel. Both values are significant to the character of this edition. Direction of travel affects the advertising content, and further study may reveal other differences in content between eastward and westward travel. The name of the ship is important because each ship had an embedded editor on board who personally composed the wireless news pages, so these pages to an extent reflect that editor's personal voice. Capturing both of these additional values for the Atlantic Edition therefore provides potential guidance to the researcher.

Assigning categories to each newspaper article is one of the most difficult tasks. This is done manually, with an operator selecting the appropriate option. The categories we use for newspapers are based on a taxonomy we have developed across many projects, with rules defining what constitutes a 'display advert', what constitutes a 'letter', what constitutes a 'news' item, and so on. As well as our standard taxonomy, we also aim to introduce categories that pull out specific features of the paper. Such unique categories are most successful when they are easy to define. In the case of the Daily Mail, we have made the 'Femail' section articles a category. A more generic 'women's page' category was much more difficult to define, as the paper went through a number of women's columns over the twentieth century, with different presentation, layouts and appearances in the paper. Knowing when to capture them was not always obvious, and if a category is captured incorrectly, its value is diluted. It was therefore decided to restrict the category to the well-known 'Femail' section instead.

XML
XML (which stands for eXtensible Markup Language) is the backbone of a digital archive. The XML files provide the structure for the various strands of data (including the OCR text and metadata), assigning tags to each element to define its role. By doing this in a clearly defined, consistent way across the whole data corpus, XML files allow a software application (i.e. the user interface) to make sense of the archive.

```
<GALENP>
<Newspaper>
<issue>
<metadatainfo>
<newspaperID>DMHA</newspaperID>
</metadatainfo>
<id>DMHA-1942-0103</id>
<is>14254</is>
<da>January 03, 1942</da>
<pf>19420103</pf>
<dw>Saturday</dw>
<ip>4</ip>
<copyright>&#x00A9; Associated Newspapers Limited</copyright>
```

```
<imdim>3784,5357</imdim>
<article type="article">
        <id>DMHA-1942-0103-0001-002</id>
                <?article zz="1860103359"?>
                        <ocr relevant="yes">43.54</ocr>
                        <sc>A</sc>
                        <pi pgref="1">DMHA-1942-0103-0001</pi>
                        <pi pgref="4">DMHA-1942-0103-0004</pi>
                        <ci pgref="1" clip="1">DMHA-1942-0103-0001-002-001</ci>
                        <ci pgref="1" clip="2">DMHA-1942-0103-0001-002-002</ci>
                        <ci pgref="1" clip="3">DMHA-1942-0103-0001-002-003</ci>
                        <ci pgref="1" clip="4">DMHA-1942-0103-0001-002-004</ci>
                        <ci pgref="4" clip="5">DMHA-1942-0103-0001-002-005</ci>
                        <pc>2</pc>
                        <ti>Vast Line-Up against Axis</ti>
                        <ta>Churchill-Roosevelt &apos;World Plan&apos; Ready</ta>
                        <ta>26 Nations Pledge &apos;No Separate Peace&apos;</ta>
                                <detailed_au>
                                        <first>Walter</first>
                                        <last>Farr</last>
                                        <suffix>Daily Mail Special Correspondent</suffix>
```

Extract of XML describing the metadata for the article "Churchill-Roosevelt 'World Plan' Ready" from 3 January 1942 issue of the Daily Mail. The XML here identifies key information about the article, such as its publication data and author, but also which image files it is associated with

```
<wd pos="547,1273,622,1298">From</wd>
<wd pos="638,1271,784,1297">WALTER</wd>
<wd pos="799,1272,890,1299">FARR</wd>
<wd pos="906,1273,989,1303">(Daily</wd>
<wd pos="1004,1273,1065,1299">Mail</wd>
<wd pos="1080,1272,1182,1303">Special</wd>
<wd pos="1199,1273,1412,1304">Correspondent)</wd>
<wd pos="1446,1286,1649,1318">Washington,</wd>
<wd pos="1668,1288,1766,1319">Friday.</wd>
<wd pos="574,1325,802,1367">BRITAIN,</wd>
<wd pos="831,1328,894,1362">the</wd>
<wd pos="925,1326,1058,1362">United</wd>
<wd pos="1087,1328,1214,1369">States,</wd>
<wd pos="1239,1328,1366,1372">China,</wd>
<wd pos="1394,1331,1529,1373">Russia,</wd>
<wd pos="1553,1333,1624,1368">and</wd>
```

Extract of XML from the same article, showing the OCR text. Note that each individual word has coordinates associated with it, which facilitates 'hit-term highlighting' in the user interface.

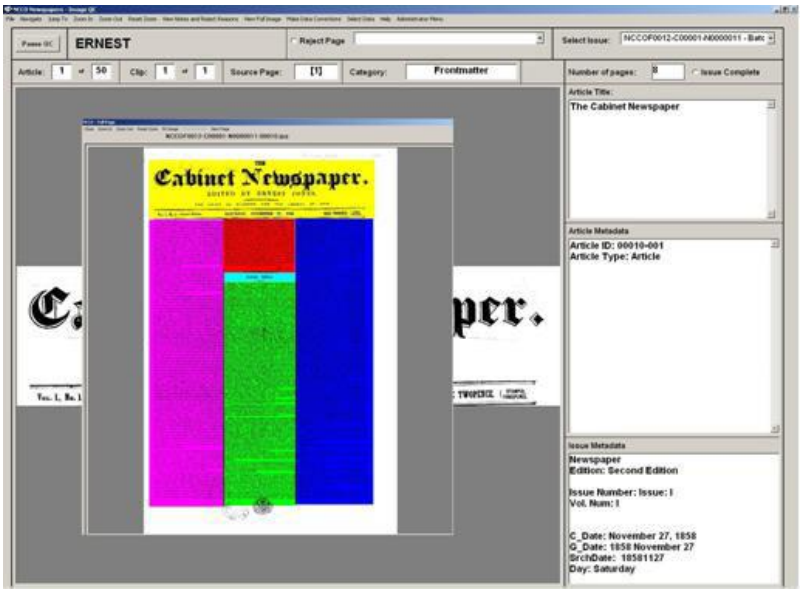
Creating XML is by far the costliest and most labour-intensive part of any digitization project. Yet, if done well, it is entirely invisible!

We start by creating a document type definition (DTD). This DTD defines the data structure for the archive, outlining a list of permissible legal elements and attributes. In essence, it provides rules and order for what would otherwise be a jumbled mass of items. All data that is captured for a project must fit the DTD, or it does not pass verification – there are no exceptions. Examples of such rules include:

There must only be one title for an article
Every article must be assigned to a specific category
Every article must have a date.

If articles fail to meet these rules, the failure is detected by our quality assurance process, and the problem can be addressed.

For newspapers, the creation of XML includes a process of 'article segmentation', whereby each individual component of a page is manually identified, and its location coordinates on the scanned image captured. This is what allows articles to be displayed as 'clips' in an archive, and also permits each article to be individually highlighted when users are looking at the page as a whole.



Example of newspaper page being segmented into individual zones

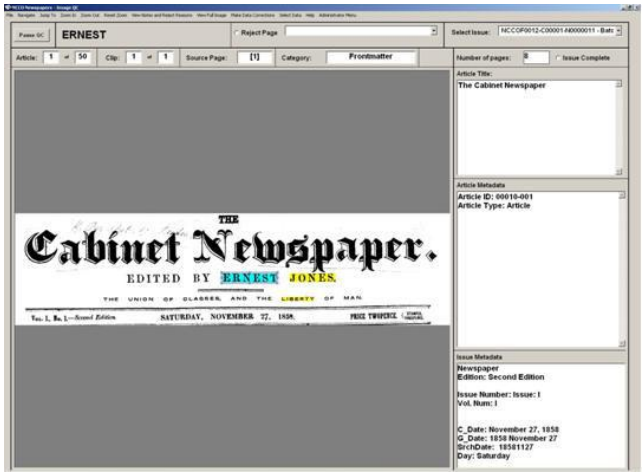
Quality Assurance

Once the XML is created, we put it through a thorough quality assurance (QA) process. Examples of these checks include:

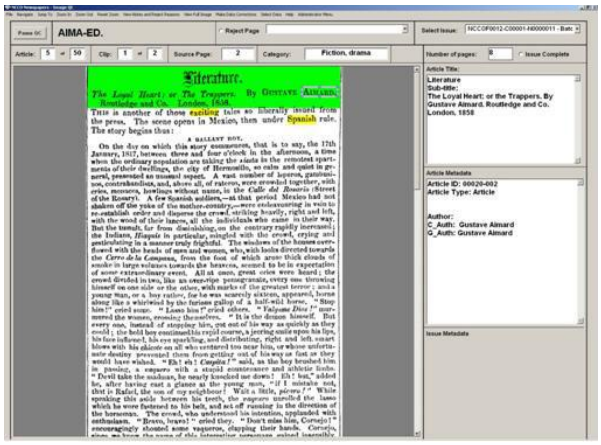
- Ensuring that an image exists for all XML references, and vice-versa
- Confirming that file naming convention and directory structure meet requirements
- Validating the XML structure against the DTD
- Checking that the image format and size meets guidelines
- Comparing the digital files against the manifest of expected items. Is anything missing or do we have more than we are expecting?

Many of the checks above can be automated, but much of our QA process is carried out by human operators, who view every page image for quality and metadata capture accuracy. This is a different team to those who create the XML, so that we can obtain an independent view of the quality of data being created.

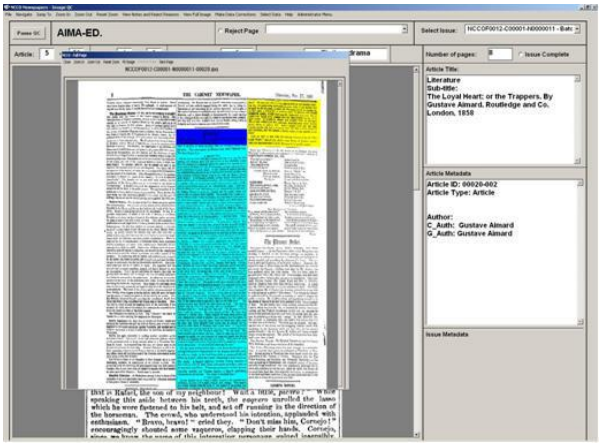
During manual QA, we will compare metadata to the source images, check the overall image quality, and confirm that the image coordinates have been captured correctly. Issues, pages and articles that do not meet our standards are rejected and returned for reworking. Those that meet acceptable standards are then moved to a staging area for the final stages of content processing, before being prepared for loading to our content delivery systems.



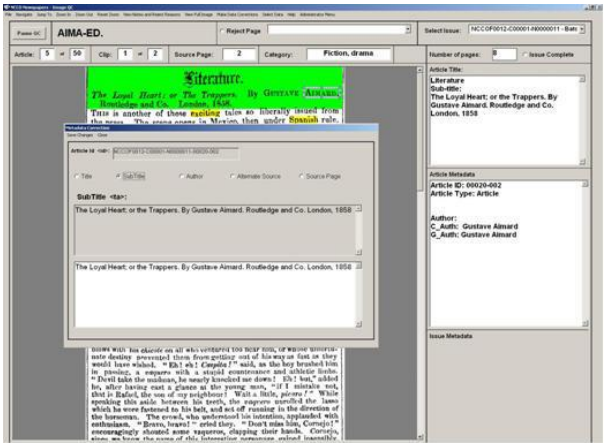
Issue level metadata is reviewed.



Article level metadata is reviewed along with zoning and article 'threading', to ensure that when an article spreads across more than one column or page that it is all captured as one article



Checking the article in the context of the full page



Operators can make data corrections while viewing each article

The Application

While we convert the content, we simultaneously create the 'application' that hosts the content. This is the front-end of the archive that users are familiar with, containing the search screens, results pages, and article display that define the way we use a digital archive. This application allows users to access the underlying database in an intuitive and straightforward manner, without the need for significant specialist knowledge.

We have an on-going process of user-testing of our products, and use this feedback to inform the development of new archives. The goal is to have a number of useful ways in which to interact with the content, without unnecessarily complicating the archive and making it inaccessible to a diverse range of users. Despite the surface-level simplicity, our newspaper archives contain a number of powerful features for advanced users. The search engine will search the OCR text, the aforementioned problems of which means that some of the retrieved results may not be appropriate, while other useful items may not have been picked up. To overcome this, the ability to use 'wildcards' as part of a search query can often help:

***** for any number of characters (e.g. **carib*** finds: Carribbean and caribou)

? in place of any one character (e.g. A search on **psych????y** finds: psychiatry and psychology but not psychotherapy)

! for one or no characters (e.g. a search on ***colo!r*** finds: color and colour)

Other powerful search tools include 'proximity operators'. These are used between two search terms to indicate that the terms must occur within a specified distance of each other. The benefit of this is that words that are close to each other are more likely to be related than words that are far apart.

A proximity operator has two components: a **letter** that indicates the *direction* and a **number** that indicates the *distance* in words. There are two proximity operators in the Daily Mail Historical Archive:

W*n* The **W** (within) operator specifies that the word that follows the operator must occur within ***n*** words *after* the word that precedes the operator. For example, the search expression **shared w3 values** matches any records in which the word *values* occurs three or fewer words after the word *shared.*

N*n* The **N** (near) operator specifies that the words on either side of the operator must occur within ***n*** words of each other *in either direction*. For example, the search expression **memory n5 repressed** matches any records in which the words *memory* and *repressed* occur within five or fewer words of each other in either direction.

Mastering the use of tools such as wildcards and proximity operators can significantly transform the experience of using a newspaper archive, and vastly improve the results retrieved.

Conclusion

Creating a digital newspaper archive is no small undertaking. Creating an archive with the careful content selection, appropriate imaging quality and richness of data required for the modern researcher is a taller task still.

Such projects are consequently a huge investment in time and resources, but by making our newspaper heritage more widely accessible and discoverable, they rescue the thoughts, words and deeds of past generations from crumbling, unnoticed, to dust. That seems a price worth paying.